Les systèmes d'intelligence artificielle pour la génération d'images

Ludovic Denoyer est professeur à Sorbonne Université, en disponibilité dans le privé.

Benjamin Piwowarski est chercheur CNRS à l'Institut des systèmes intelligents et de robotique (Sorbonne Université).

Les systèmes de génération d'images (Midjourney, Stable Diffusion ou Dall-E pour ne citer que certains des plus connus) font beaucoup parler d'eux du fait de la qualité des résultats produits. Nous avons tous pu voir un Donald Trump arrêté par des policiers, un Emmanuel Macron en gilet jaune, un pape habillé avec une doudoune… Le réalisme est saisissant. Ces images ne sont pas issues d'un montage Photoshop, mais bien le résultat du travail de systèmes d'intelligence artificielle. Certaines images ainsi générées ont même été primées lors de concours photo, sans que le jury ne détecte une « supercherie » [1]. Comment des systèmes informatiques arrivent-ils à générer des images inédites d'une grande qualité qui les rendent parfois indiscernables d'une photo réelle ? Nous proposons ici une description des principes de leur fonctionnement.

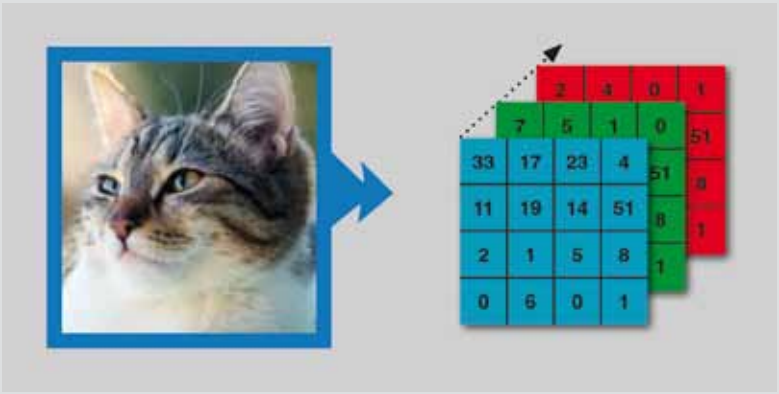
Deux études de la tête d'un vieillard, Jacob Jordaens (1593-1678)

Représentation sémantique

Un défi majeur de la recherche en intelligence artificielle réside dans la définition d'un espace sémantique pour décrire les images. En effet, dans un ordinateur, une image est représentée comme une suite de valeurs numériques correspondant à l'intensité lumineuse et à la couleur des pixels la constituant. Ainsi, une image peut être représentée par une sorte de tableau de valeurs numériques de très grande dimension (on parle de « tenseur de valeurs numériques » ; voir l'encadré « Représentation d'une image par pixels »).

Manipuler des structures de cette nature est compliqué pour les besoins qui sont ceux de la génération d'images. Le problème réside dans le fait que de telles représentations ne sont porteuses d'aucune sémantique : ils codent la manière dont l'image doit s'afficher à l'écran, mais ne disent rien de son contenu. Ainsi, prenons l'exemple de deux photographies d'une même scène, mais prises sous deux angles légèrement différents. Bien que les contenus soient sémantiquement proches (et donc que les images soient interprétées comme similaires par un

Représentation d'une image par pixels



Une image correspond à un « tenseur » en trois dimensions qui encode la couleur de chaque pixel. Supposons, pour simplifier, que le chat soit représenté par 16 points (pixels). Le pixel en haut à gauche sera alors décrit par l'intensité des couleurs correspondantes (ici : 33 pour le bleu, 7 pour le vert et 2 pour le rouge – la combinaison de ces trois couleurs de base formant la couleur du pixel). En réalité, pour bien représenter un chat, il faut une résolution bien supérieure à 16 pixels. Les appareils photos de nos téléphones portables ont maintenant des résolutions de quelques mégapixels. Ceci signifie que le tableau, au lieu d'être de 4 sur 4 comme sur la figure, sera de 1 000 sur 1 000 pour un mégapixel (donc trois millions de valeurs pour représenter l'image).

être humain), les représentations associées seront considérablement différentes si l'on compare les pixels correspondants. Il est donc nécessaire de construire un espace de représentation des images plus pertinent – et également facile à manipuler : c'est *l'espace sémantique*.

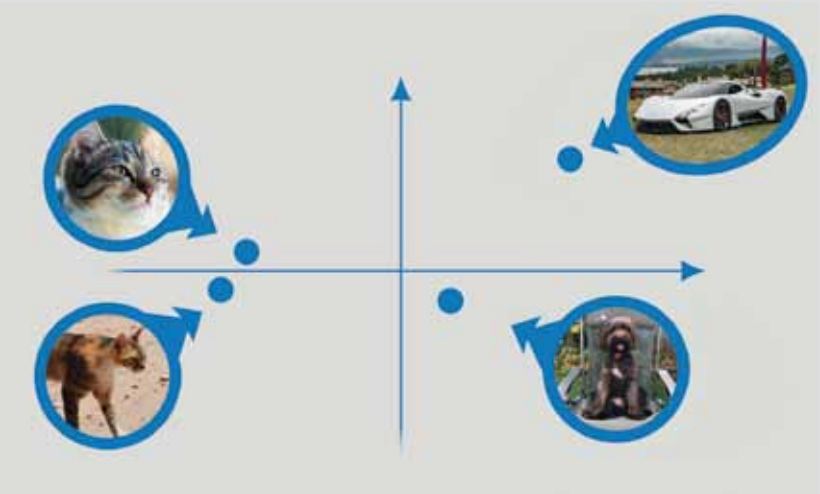
L'espace sémantique est basé sur un nombre de dimensions réduites, généralement de l'ordre du millier, à comparer aux millions de dimensions utilisées pour décrire une image sous forme de pixels. Dans cet espace, chaque image est représentée par un point. Le principe fondamental qui sous-tend la création d'un tel espace est le suivant : tandis que deux images de sémantique proche (par exemple, deux images du même chat sous un angle distinct, ou à des instants séparés) peuvent être très différentes dans l'espace des pixels, l'espace sémantique est capable de capturer cette similarité sémantique. Ainsi, ces deux images seront représentées par deux points rapprochés dans cet espace (voir l'encadré « L'espace sémantique »). Dans l'espace sémantique, l'information contenue dans l'image est comprimée en un nombre réduit de paramètres. Si, dans leur codage initial, les dimensions correspondent à la couleur de chaque pixel, l'espace sémantique va, lui, capter des éléments tels que le type d'objet, la luminosité, l'orientation de l'objet représenté, etc. Il permet donc de représenter les images avec une information bien plus pertinente que celle véhiculée par chaque pixel individuel. En revanche, il ne permet pas d'afficher l'image à l'écran : il s'agit d'une représentation abstraite. Il sera nécessaire, nous allons le voir, d'assurer un lien entre les deux types de représentation.

Comment construire l'espace sémantique ?

Cet espace ne peut pas être construit manuellement, et c'est là que les méthodes d'intelligence artificielle (IA) interviennent. Pour qu'un ordinateur puisse construire et exploiter un tel

espace, il est essentiel de savoir comment transformer (ou projeter) une image dans cet espace, et inversement, comment reconstituer (ou extraire) une image à partir d'un point dans cet espace. La première opération est désignée sous le terme d'encodage tandis que la seconde est appelée décodage. Ces deux opérations sont effectuées par des réseaux de neurones spécifiques appelés « encodeur » et « décodeur » (voir encadré « Encodage et décodage sémantiques d'une image »).

L'espace sémantique



Un espace sémantique en deux dimensions. Chaque image correspond à un point dans cet espace, et les positions relatives dans l'espace capturent une similarité sémantique entre images : par exemple les représentations sémantiques des deux images de chats sont proches, mais éloignées de la représentation sémantique d'une voiture. En réalité, les espaces sémantiques ont bien plus de deux dimensions (de l'ordre du millier).

Encodage et décodage sémantiques d'une image

L'encodeur transforme une image sous forme de pixels en un point dans l'espace sémantique. Le décodeur transforme un point de l'espace sémantique en une image sous forme de pixels. Ces deux composants sont réalisés par des réseaux de neurones « entraînés » pour cette finalité.

La question qui se pose alors est : comment construire cet encodeur et ce décodeur ? Pour découvrir (« apprendre ») ces deux fonctions, divers algorithmes d'intelligence artificielle ont été proposés. Ils utilisent un ensemble d'images (appelé ensemble d'apprentissage) afin de construire l'espace sémantique. À noter que des images brutes (sans aucune annotation) sont les seuls ingrédients nécessaires ici. L'approche standard consiste à utiliser une méthode dite d'« auto-encodage ».

Le principe est le suivant. En supposant que le décodeur et l'encodeur sont parfaits, et que l'espace sémantique contient suffisamment d'information pour reconstituer n'importe quelle image, si je prends une image, que je l'encode comme un point de l'espace sémantique, puis que je décode ce même point, on devrait retrouver l'image initiale. En pratique, à travers des itérations successives (opération appelée la « descente du gradient »), l'algorithme va progressivement définir un espace sémantique, et mettre au point un encodeur et un décodeur appropriés. Ainsi, dans ce système, l'algorithme d'IA vise à apprendre à la fois à l'encodeur et au décodeur à reconstituer chaque image, pour un ensemble d'images donné (l'ensemble d'apprentissage), qui soit aussi proche que possible de l'image initiale.

L'idée clef derrière ce processus est que, l'espace sémantique étant plus petit que l'espace des pixels, l'encodeur va agir comme un compresseur de données et être forcé à découvrir des dimensions qui encodent le contenu informationnel de l'image de manière pertinente. Ainsi, par exemple, là où la présence d'un chat dans l'image est encodée à travers de multiples dimensions dans l'espace des pixels, l'espace sémantique pourra capturer sur une ou quelques dimensions le type de chat représenté dans l'image, et le décodeur saura retraduire cette information sous forme de pixels.

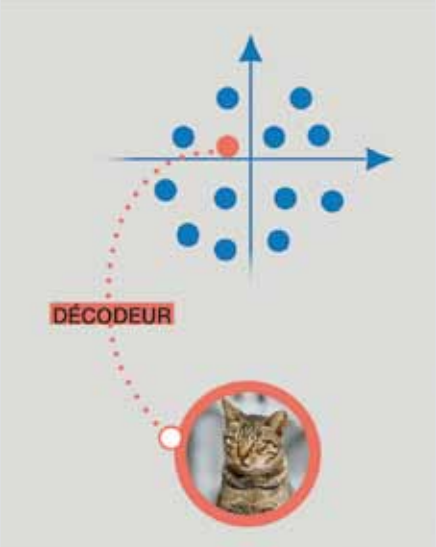
Génération d'images

Maintenant que nous avons exposé les principes de l'apprentissage d'un espace sémantique, comment se déroule concrètement la génération de nouvelles images ?

Une fois qu'un espace sémantique a été construit à partir d'un ensemble d'images, la projection, *via* l'encodeur, de toutes ces images dans cet espace crée un nuage de points (les points bleus dans l'encadré « Génération d'images à partir d'un point de l'espace sémantique »). Si l'on choisit aléatoirement l'un de ces points et que l'on reconstitue l'image correspondante à l'aide du

Génération d'images à partir d'un point de l'espace sémantique

L'espace sémantique contient les représentations sémantiques (points bleus) des images d'apprentissage utilisées pour instruire l'encodeur et le décodeur. Si l'on choisit un point au hasard dans cet espace et qu'on le décode, alors une nouvelle image est générée, sémantiquement proche des points bleus voisins ayant servi lors de l'apprentissage.



décodeur, alors nous allons reconstituer une image aléatoirement choisie de notre ensemble original, respectant en cela le principe de l'auto-encodage. Le principe d'apprentissage utilisé permet d'obtenir une restitution potentiellement parfaite.

Mais qu'arrive-t-il si je sélectionne un point aléatoirement dans tout l'espace sémantique, c'est-à-dire non plus un des points qui a été généré lors de la phase d'apprentissage (les points bleus dans l'encadré), mais un point au hasard (par exemple le point orange dans l'encadré) ? Dans ce cas, le décodeur va construire une image originale, différente de celles de l'ensemble d'apprentissage, mais proche sémantiquement des points voisins. Par conséquent, générer une nouvelle image consiste à sélectionner aléatoirement un point dans l'espace sémantique et à le décoder. C'est le principe fondamental des modèles génératifs d'images.

Mais une question demeure : comment choisir la région de l'espace sémantique qui permette de générer une image réaliste ?

Générer des images réalistes et de qualité

Si le point sélectionné aléatoirement se situe « à l'intérieur » d'un nuage de points formés lors de la phase d'entraînement et de construction de l'espace sémantique, il est fort probable que l'image reconstituée ressemble aux images de l'ensemble d'apprentissage à l'origine de ces points, donnant ainsi une image « réaliste ». Pour simplifier et illustrer, si le « point orange » de l'encadré est choisi près des points bleus qui représentent le concept de chat, il est probable que l'image générale sera un chat très réaliste. Par contre, choisir un point au hasard n'importe où dans l'espace sémantique risque d'aboutir à une image qui ne corresponde à rien.

Ainsi, les méthodes de génération vont principalement ajouter aux techniques d'encodage décrites précédemment des systèmes permettant d'assurer que les images générées proviennent d'un endroit pertinent de l'espace sémantique. Le réalisme des images ainsi généré a provoqué un véritable émoi dans la communauté. La qualité s'est ensuite progressivement améliorée pour atteindre le niveau que l'on connaît aujourd'hui (voir l'encadré « Auto-encodeurs variationnels et modèles de diffusion »).

Nature morte aux oranges, Louis Hayet (1864-1940)

Générer des images à partir de texte

Les modèles décrits précédemment sont capables de générer des images, mais sans que nous puissions en spécifier le contenu. La question qui se pose alors est comment permettre à un utilisateur de piloter la génération, notamment en écrivant un texte (appelé « prompt ») décrivant le type d'image qu'il souhaite produire.

Nous avons expliqué la construction d'un espace sémantique pour les images grâce à l'utilisation d'« auto-encodeurs ». Les mêmes principes vont être appliqués aux données textuelles (voir l'article sur les modèles de langue dans ce dossier). Un espace sémantique pour le texte de la requête va ainsi être construit. Nous disposons alors de deux espaces sémantiques : l'un pour les images et l'autre pour le texte. Reste à être capable de passer de l'un à l'autre pour générer les images à partir d'une description textuelle.

Une manière de procéder consiste à coder le texte dans le même espace sémantique que les images. Il sera alors possible de décoder ce texte sous forme d'image (à l'aide d'un modèle d'auto-encodage décrit précédemment). Les images utilisées pour qu'un algorithme d'IA apprenne cette correspondance sont étiquetées avec des textes descriptifs de leur contenu. Un espace sémantique commun aux images et aux textes est ainsi construit par apprentissage. Les modèles

« Auto-encodeurs variationnels » et « modèles de diffusion »

Les « auto-encodeurs variationnels » pour des images réalistes

Le principe des auto-encodeurs variationnels [1] repose sur l'ajout d'une contrainte au processus d'auto-encodage pour faire en sorte que la distribution des points dans l'espace sémantique corresponde à une distribution connue appelée « bruit ». Dit autrement, ce modèle s'assure que les images de notre base d'entraînement correspondent à du « bruit » dans l'espace sémantique, et donc que, si l'on génère du bruit que l'on décode en image, alors l'image résultante sera réaliste.

L'usage des auto-encodeurs variationnels pour la création d'images a permis la génération d'images d'un réalisme saisissant. Toutefois, ces images sont loin d'être parfaites et leur qualité est insuffisante pour tromper l'œil humain. Cette imperfection s'explique principalement par le fait que les fonctions associées susceptibles de projeter depuis l'espace des images vers l'espace sémantique (encodeur) ainsi que le décodeur sont, *a priori*, des fonctions potentiellement très complexes et difficiles à apprendre à partir de données.

Les modèles de diffusion pour améliorer la qualité des images

Les modèles de diffusion [2] vont venir améliorer la situation. Ils reposent sur le principe que l'on peut apprendre à passer d'une distribution *artificielle* des données (le bruit) à une distribution *naturelle* (les pixels) en procédant par étapes successives. Ainsi, les modèles de diffusion sont des modèles dont l'encodeur bruite une image petit à petit et le décodeur opère un débruitage séquentiel (de l'ordre d'une centaine d'étapes) pour enlever progressivement le bruit dans les images, comme illustré dans la figure. Si l'on génère un bruit aléatoire, alors le décodeur va générer une image nouvelle.

L'apprentissage d'un modèle de diffusion sur un ensemble d'images consiste donc à entraîner un réseau de neurones à réduire le bruit des images préalablement bruitées. Travailler dans l'espace des images est complexe en termes de calcul car il faut travailler dans un espace composé de millions de dimensions. Cependant, il est possible de travailler directement dans l'espace sémantique précédemment décrit qui est de plus petite dimension. Les modèles actuels, de façon très grossière, reposent sur le principe combiné de l'apprentissage d'un auto-encodeur et d'un modèle de diffusion *dans l'espace sémantique.*

En résumé, un modèle de diffusion n'est rien de plus qu'un modèle de débruitage permettant de reconstruire les images d'un ensemble de données fourni par un utilisateur. Si cet ensemble de données contient des images de chats et de chiens, le modèle générera des images de chats et de chiens. Mais si l'ensemble d'apprentissage est constitué de millions d'images extraites du Web, alors le modèle apprendra à générer des images de

tous styles. C'est le cas par exemple avec le logiciel Stable Diffusion, qui est une version open source de ce type de modèle (le code informatique du logiciel est librement accessible, il peut être consulté et modifié).

Références

[1] Diederik P et al., "Auto-encoding variational bayes", International Conference on Learning Representations, 2014. Sur arxiv.org

[2] Song Y et al., "Score-based generative modelling through stochastic differential equations", 2021. Sur arxiv.org

apprennent ensuite une fonction permettant de passer de l'espace sémantique du texte vers l'espace sémantique des images. Une fois cette projection apprise, il devient possible de transformer du texte en une image générée : le texte est encodé dans l'espace sémantique textuel, puis « traduit » dans l'espace sémantique des images, avant d'être décodé en image.

L'approche adoptée par les modèles disponibles actuellement sur Internet, tels que Stable Diffusion [2] ou Midjourney [3], est un peu plus complexe car elle repose sur les modèles de diffusion et non d'auto-encodage : dans ces systèmes, on utilise l'espace sémantique du texte pour « guider » la génération d'une image. Chaque étape du débruitage est influencée par le texte, c'est-à-dire par sa représentation dans l'espace sémantique textuel.

Même si ces deux possibilités de générer une image à partir d'un texte semblent très différentes, elles reposent toutefois sur des principes très similaires, car ancrées dans les espaces sémantiques appris automatiquement par des algorithmes d'IA.

Comme pour la génération « libre » d'image, le succès de ces méthodes provient de la capacité à entraîner ces modèles sur de très grandes quantités de données directement extraites du Web.

Conclusion

Tout comme pour la génération de textes, les principes clés des modèles génératifs d'images sont finalement assez simples, même si les concepts mathématiques sous-jacents peuvent être compliqués. Toutefois, alors que les modèles textuels ont surtout progressé grâce à l'émergence de nouvelles architectures de réseaux de neurones et à l'utilisation de principes d'apprentissage éprouvés, les modèles génératifs d'images ont, quant à eux, permis l'émergence de nouvelles familles de méthodes, comme les « auto-encodeurs variationnels », les « modèles de diffusion », ou encore les « modèles génératifs adversariaux » dont nous n'avons pas discuté ici. Alimentés par des quantités de données astronomiques lors de la phase d'apprentissage, ces modèles produisent aujourd'hui des images pour lesquelles il devient très difficile, voire impossible, de distinguer le caractère artificiel.

Cependant, ces modèles ont encore des limites. La principale réside dans la quantité de données (et de calculs) nécessaire à leur apprentissage, ainsi que le coût sous-jacent (en temps de calcul). De nombreuses recherches actuelles se concentrent sur la réduction des ressources machines requises, et nous voyons déjà de grands progrès dans cette direction. Cela laisse à penser que, dans quelques années, ce type de modèle sera accessible à tous. La deuxième limite est davantage sociétale que technologique : si des images indiscernables d'images réelles peuvent être générées, quels en sont les dangers ? Cette question, qui a été largement négligée par la communauté scientifique et politique, doit être abordée rapidement. Cependant, elle dépasse largement le cadre de cet article. //

Ludovic Denoyer et Benjamin Piwowarski

Références

[1] Gayte A, « L'image d'une IA a dupé les organisateurs du plus prestigieux concours de photos », *Numerama*, avril 2023. Sur numerama.com

[2] Site de Stable Diffusion. Sur stablediffusion.com

[3] Site de Midjourney. Sur midjourney.com